Integrating Value Function-Based and Policy Search Methods for Sequential Decision Making

(Extended Abstract)

Shivaram Kalyanakrishnan
Department of Computer Sciences, The University of Texas at Austin
shivaram@cs.utexas.edu

Keywords

Reinforcement learning, Temporal difference learning, Policy search, Function approximation.

1. THESIS TOPIC

Sequential decision making from experience, or reinforcement learning (RL), is perfectly suited to autonomous agents that are situated in an unknown environment, and which must use their interactions with the environment to learn behavior that maximizes long-term gains. In general, this setting can be treated as a Markov Decision Problem (MDP), comprising a set of states S, a set of actions A, a reward function $R : S \times A \times S \rightarrow \mathbb{R}$, and a transition function $T : S \times A \times S \rightarrow [0, 1]$. In an MDP, the objective is to find a policy $\pi : S \rightarrow A$ that maximizes the expected long-term reward from every state $s \in S$. This can be done by determining the optimal action value function $Q^* : S \times A \rightarrow \mathbb{R}$, from which the optimal policy, denoted π^*, can be derived as: $\pi^*(s) = \text{argmax}_a \, Q^*(s, a), \forall s \in S$.

Classical approaches such as temporal difference learning [6], which proceed by successively refining the action value function based on observed experiences, provide efficient solutions to MDPs with finite sets of states and actions. Yet, a predominant number of sequential decision making problems that arise *in practice* have continuous (or very large) state spaces, which force the use of function approximation. Further, in many applications, sensor noise corrupts the state signal. As a consequence, nearly every RL problem in practice corresponds to a Partially Observable MDP (POMDP), to which most of the theoretical guarantees of value function-based (VF) methods fail to extend. Coping with partial observability in a principled manner has merited considerable attention in the literature [2], but is yet to scale to complex tasks with continuous state spaces.

Policy Search (PS) methods [1, 7] are optimization methods that directly seek to find parameters $\mathbf{w}^*$ of the optimal policy π^* by searching through the space of parameters W. In so doing, they do not necessarily compute the value function of the policy, and consequently, are likely to be less sample-efficient than VF methods. At the same time, their asymptotic performance is likely to be affected less by function approximation and partial observability. For most sequential decision making problems that arise in practice, there exist no theoretical bounds for the sample efficiency or asymptotic performance of either VF or PS methods; it is left to empirical devices to ascertain how these contrasting method perform.

This thesis aims to develop learning methods for practical sequential decision making tasks by integrating VF and PS methods, with the objective of achieving both sample efficiency and superior asymptotic performance.

2. COMPLETED WORK

2.1 Empirical Analysis of VF and PS Methods

As the first step towards combining the merits of VF and PS methods, we conduct a systematic empirical study to examine their relative strengths and weaknesses [3]. To do so, we devise a suite of "grid world" domains that can be varied for four parameters: problem size s, action noise p, expressiveness of function approximation χ, and state noise σ. Across a broad range of parameters settings (1250 in total), we record the performance of Sarsa, a classical VF method, and cross entropy optimization, a PS method.

We see clear patterns in the domain characteristics for which each class of methods excels. Our experiments illustrate that VF methods enjoy superior sample complexity and asymptotic performance when provided precise function approximators and complete state information. However, with inadequate function approximation and noisy state information, their performance drops significantly, and indeed below the asymptotic performance achieved by PS methods. With fixed values of $s = 10$, $p = 0.3$, and $\sigma = 0$, we observe the effect of varying the function approximation paramater χ in Figures 1(a) and 1(b). At $\chi = 1$ (exact representation of state space), VF indeed converges to the optimal policy, and at a much quicker rate than PS. Yet, under a deficient representation ($\chi = 0.1$), VF performs very poorly when compared to PS, which does not show such a drastic drop in asymptotic performance. Increasing the state noise σ adversely affects the asymptotic performance of both VF and PS methods, although the decline is more gradual for PS.

We implement a simple scheme to integrate VF and PS, which we enforce to share the same representation. In this integrated method, VF+PS, the learned representation of VF after a certain number of episodes of learning is transferred to PS. As visible in Figures 1(a) and 1(b), VF+PS inherits both the superior sample efficiency of VS and the high asymptotic performance of PS. Not only does VF+PS achieve higher asymptotic performance than both VF and

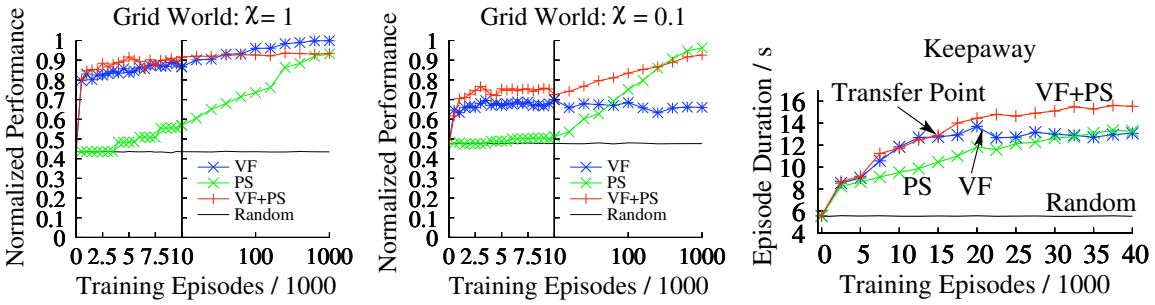

Figure 1: Empirical Analysis of VF and PS methods. In (a) and (b), note the break in the x axis at 10,000 episodes, beyond which a log scale is adopted. Descriptions are provided in text.

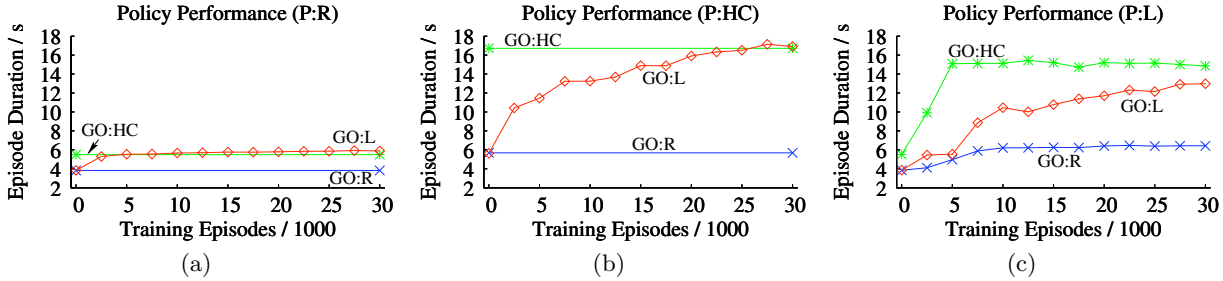

Figure 2: Keepaway Pass and GetOpen. The graphs shows three GetOpen policies (Random (GO:R), Hand-coded (GO:HC) and learned (GO:L)) when paired with a Pass policy that is Random (P:R (a)), Hand-coded (P:HC (b)), or Learned (P:L (c)).

PS on a majority of our test settings, we also demonstrate its effectiveness on the more complex Keepaway task in robot soccer [5] (Figure 1(c)).

2.2 VF+PS for a Complex Multiagent Task

Whereas previous successful results in the Keepaway task have limited learning to an isolated, infrequent decision that amounts to a turn-taking behavior among players (Pass), we expand the agents' learning capability to include the more ubiquitous action of moving without the ball (GetOpen) [4]. GetOpen induces a complex MDP, which is not suitable to be learned by VF approaches, such as the one employed by Stone *et al.* for learning Pass. Unlike Pass, there are multiple players executing GetOpen at any instant of time. We provide a PS method for learning GetOpen. As a result, we learn a composite behavior (Pass+GetOpen) in which multiple agents execute *learned* policies simultaneously.

As reported in Figure 2, the learned GetOpen policy (GO:L) matches the best hand-coded policy for this task (GO:HC) when paired with a hand-coded Pass policy (P:HC). Indeed GO:L outperforms GO:HC when paired with a random Pass policy (P:R). Importantly, we notice that Pass and GetOpen can be learned simultaneously, signifying that a very complex multiagent task can be completely learned by decomposing it into components that are learned separately by VF and PS methods (Figure 2(c)).

3. PROPOSED WORK

In our empirical analysis, we identify three relevant classes of methods to include in our study: actor-critic algorithms, policy gradient methods, and VF methods using eligibility traces [3]. All these methods show some degree of resistance to deficient function approximation and partial observability; we aim to include them in our comparison of VF and PS methods. Intelligently determining the "transfer point" in our VF+PS algorithm, i.e., when to stop applying VF and switch to PS, constitutes yet another problem for proposed research.

One of the reasons PS methods such as evolutionary algorithms are not sample-efficient is because they have to negate the stochasticity in fitness estimates of candidate solutions by taking an average over multiple evaluations. Currently we are currently working on a statistical technique to reduce the number of such evaluations needed to get reliable estimates. Needless to say, we seek to extend our results from the Keepaway domain to other complex, realistic sequential decision making tasks.

4. REFERENCES

[1] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. Incremental natural actor-critic algorithms. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 105–112. MIT Press, Cambridge, MA, 2008.

[2] A. R. Cassandra, L. P. Kaelbling, and M. L. Littman. Acting optimally in partially observable stochastic domains. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume 2, pages 1023–1028, Seattle, Washington, USA, 1994. AAAI Press/MIT Press.

[3] S. Kalyanakrishnan and P. Stone. An empirical analysis of value function-based and policy search reinforcement learning. In *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems*, May 2009. To appear.

[4] S. Kalyanakrishnan and P. Stone. Learning complementary multiagent behaviors: A case study. In *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems*, May 2009. To appear as short paper.

[5] P. Stone, R. S. Sutton, and G. Kuhlmann. Reinforcement learning for RoboCup-soccer keepaway. *Adaptive Behavior*, 13(3):165–188, 2005.

[6] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[7] S. Whiteson and P. Stone. Evolutionary function approximation for reinforcement learning. *Journal of Machine Learning Research*, 7:877–917, May 2006.